Assignment 05

(5 points)

In this assignment, you use the bankruptcy data set from a paper (see the bottom of the assignment). The data set can be downloaded from HuskyCT. The following is the information of the attributes, from the income statements and balance sheets:

1) Size
    a. Sales
2) Profit
    a. ROCE: profit before tax=capital employed (%)
    b. FFTL: funds flow (earnings before interest, tax & depreciation)=total liabilities
3) Gearing
    a. GEAR: (current liabilities + long-term debt)=total assets
    b. CLTA: current liabilities=total assets
4) Liquidity
    a. CACL: current assets=current liabilities
    b. QACL: (current assets – stock)=current liabilities
    c. WCTA: (current assets – current liabilities)=total assets
5) LAG: number of days between account year end and the date the annual report and accounts were filed at company registry.
6) AGE: number of years the company has been operating since incorporation date.
7) CHAUD: coded 1 if changed auditor in previous three years, 0 otherwise
8) BIG6: coded 1 if company auditor is a Big6 auditor, 0 otherwise

The target variable is FAIL, either = 1 or 0. You program and model using logistic regression.

In the main code, you can see instead of gradient descent, a MATLAB function fminunc is used to search for the optimum (fminunc is for unconstrained optimization, and fmincon is for constrained optimization). Data is normalized. The first part is without regularization and will give a reasonable accuracy (~80%) with the whole data set considered as the training data set. In the second part, regularization is utilized. The function mapping maps variables to a higher dimensional space. For example from

$$\begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

to

$$\begin{bmatrix} 1 \\ x_1 \\ x_2 \\ {x_1}^2 \\ x_1 x_2 \\ {x_2}^2 \end{bmatrix}$$

to fit the data better. Then fminunc is called to search for the optimum. All the data are used to predict (~100%). Here the mapping and regularization improve the performance.

1. (4 points) Read, understand, and run the MATLAB code.

Pick either R or Python to do the work based on the working MATLAB code is given to you. 1 additional point if you write code in both Python, and R. Logics, variable names, and function names should follow the ones in the given MATLAB code as much as possible. In your code, instead of using all data, randomly sample from the dataset, take 60% for training, and use the rest of observations for testing the algorithm.

In Python, you can use the minimize function in the scipy.optimize package:

http://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html

There are several optimizations methods by the method argument. You can use any, or can use your own Gradient Decent.

In R, use the optim function in the optim package

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/optim.html

However, there is no restriction on what package you can use

2. (1 point) Explain your code ***in details***, show your understanding (you can do this line by line, or small blocks of code).

Submission: your script, and a document explaining your work.

1. Malcolm J. Beynon, Michael J. Peel, Variable precision rough set theory and data discretisation: an application to corporate failure prediction